Examining the Effect of a Game-like Practice Tool on the Quality of Student Peer Evaluations

Daniel M. Ferguson
Engineering Education
Purdue University
West Lafayette, IN USA
dfergus@purdue.edu

Elizabeth Shu
Multidisciplinary Engineering
Purdue University
West Lafayette, IN USA
eshu@purdue.edu

Yuchen Cao
Statistics
Purdue University
West Lafayette, IN USA
cao273@purdue.edu

Matthew Ohland
Engineering Education
Purdue University
West Lafayette, IN USA
ohland@purdue.edu

***Abstract* — This work-in-progress research-to-practice analysis examines whether using a more game-like practice tool for learning a peer rating schema significantly improves the quality of peer evaluations in a team-based course. Peer evaluations in our study are defined as the peer evaluations of teamwork behavior as measured by the CATME peer evaluation system. Data samples in this study include students in teams from introduction to engineering courses in a large Midwest university's engineering program who used the previous version and the current more game-like version of the rater practice tool. We found significant differences in the mean ratings of our intervention group and in the distribution of variances versus our control group in the first peer evaluation of the semester.**

Keywords— teamwork, peer ratings

I. Introduction

Teamwork behaviors are key attributes sought after by a of companies when hiring new employees. In addition, they assert that working in teams not only helps distribute the workload better but leads to greater efficiency, better communications in the future and creates an environment for workers that can serve as a platform for even better performance [1, 2]. Hence, teamwork skills training has become more prevalent throughout college programs and in businesses [3]. Teamwork is defined as "a cooperative or coordinated effort on the part of a group of persons acting together" [4]. Chen argues that many students entering the workplace lack key teamwork skills that hamper their abilities to excel in their job field [5]. Our research question was:

Is the quality of peer evaluations significantly improved by use of a more game-like Rater Practice tool?

CATME (Comprehensive Assessment of Team Member Effectiveness) is a web based tool in use by over 8,000 instructors, across multiple disciplines, in over 2,000 institutions worldwide was created for the academic teamwork environment. CATME is used to form teams and to assist students in giving peer feedback to their team members [6]. Prior research has shown that the accountability of ratings given to or by an individual play a huge role in reflecting the accuracy of the rating and this feature is an integral part of CATME [7, 8]. CATME is constructed around five behavioral dimensions: Having Relevant Knowledge, Skills and Attributes (KSAs), Contributing to the Team's Work, Interacting With Teammates, Keeping the Team on Track, and Expecting Quality [9, 10]. These dimensions are defined as follows:

Having (H) relevant KSAs refers to the base knowledge of individual team members.

Contributing (C) to the Team's Work is being able to add value to your team's work/project.

Interacting (I) with teammates refers to the way individuals communicate within their teams.

Keeping (K) the team on track is similar to being a timekeeper.

Expecting (E) quality is taking expectations to the next level and working collaboratively to produce the best possible team outcomes. [9].

Peer evaluations facilitate better learning outcomes in upper level education, encouraging students to continue their engagement with constructive team behavior in future team activities [11]. A peer evaluation is an assessment of an individual's contribution to a work activity by their peers [4]. Peer evaluations help individuals and teams learn how to act and assess one another's performance. Peer evaluations, as facilitated by the CATME system, potentially point out behaviors where an individual excels and areas where he/she may need improvement [12, 13].

II. Research Methods

Measuring the quality of peer evaluations

The quality of a peer rating in this paper is first defined as having a larger dispersion (measured by standard deviation) in the ratings of teamwork dimensions [14]. A larger dispersion shows that an individual is differentiating their perception of a teammate's contribution on each particular CATME teamwork dimension [15]. While it is clearly possible that an individual exhibits the same level of performance across all 5 CATME dimensions, it is not likely, and similarly it is unlikely that there are no differences in teamwork performance among all team members on any of the 5 CATME dimensions [14, 16].

978-1-5386-1174-6/18/$31.00 ©2018 IEEE

Second we examined the variances in self and peer ratings using the Social Relations Model (SRM) [17-19], SRM identifies three different variance components:

1. how individuals perceive others (Rater effect), **Rater Effect** measures how consistently an individual rates his/her teammates. A **larger** Rater Effect would mean that on average an individual gives all his/her teammates the same rating. (unreliable rating)
2. how individuals are perceived by others (Target Effect), **Target Effect** measures how consistently an individual is rated by his/her teammates. A **larger** Target Effect would mean that on average an individual receives similar ratings from all his/her teammates. (reliable rating)
3. the extent to which the ratings given and received by each pair are related to each other (Relationship Effect). **Relationship Effect** measures the uniqueness of relationships after the rater and target variance have already been accounted for. A **larger** Relationship Effect would indicate that ratings are based on personal interaction inside or outside of team activities that may not be experienced by others on the team.

Feedback from Peer Evaluations and Rater Practice (RP)

CATME peer evaluations are conducted online and students receive feedback in an online form as pointers to verbal descriptions of behaviors of their self-rating, the average rating from their peers, and the average rating for all team members as shown in Figure 1.

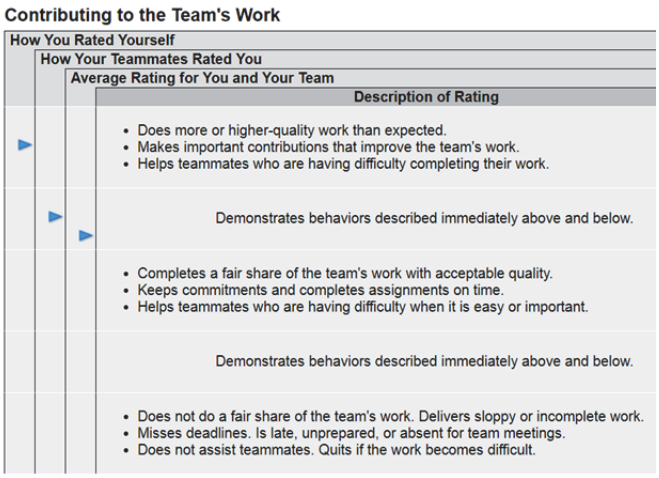

Figure 1: Sample feedback screen from the CATME Peer Evaluation tool

Although the behavioral anchors are converted to numbers in the instructor interface, numbers are omitted in student feedback so that students can continue to focus on behavior. Averaged ratings are interpreted as pointer/arrow placements on feedback screens because the ratings from their teammates are kept confidential [13]. Rater Practice training uses the same display formats for providing student feedback. The first version of Rater Practice was updated to Rater Practice 2.0 to have only 3 students not 4 students and there were no other major changes RP 1.0 to RP 2.0

Redesign of the Rater Practice tool

For the redesign of the rater practice tool, we conducted multiple structured interviews with instructors requiring rater practice in their courses and had those instructors complete RP themselves and watch a tutorial video about RP. We also observed students doing the rater practice exercise in both the old and new formats. These students included individuals who had never used CATME. We interviewed students during and after they conducted RP. From our interviews, game design research [20], and observation we learned:

1. Feedback on RP came at the end of the RP 2.0 exercise, which was frustrating to students, and the rater practice exercise was judged boring.
2. Students in classes using CATME did RP 2.0 without explanation or training on use of the rating schema.
3. Students did RP 2.0 only if it was required and then as quickly as possible.

From a review of game design and simulations to encourage learning we learned that:

1. There should be no winners and losers
2. There should be active learner control and rapid feedback.
3. The user should have a goal orientation.
4. The user should not be competing.
5. The user should have control of the game/simulation [21].

And overall from this review of prior art and game simulation research we concluded that simulation of peer ratings was our best learning goal in order to facilitate peer evaluation learning [19, 22].

Based on this feedback, we revised the design of RP to create the Rater Practice 3.0 tool in a variety of ways:

1. Individual RP average and high scores and the class high score are displayed on the student's results page.
2. RP is now associated only with a specific peer evaluation survey so that it can be assigned like class homework.
3. RP can be replayed with different student scenarios as often as the student elects.
4. Students are given immediate feedback after scoring each CATME dimension.
5. The RP scoring system is similar to that of horseshoes: A full point if correct, a half-point if 1 category off.
6. An instructor Rater Practice package was developed with three 'how to' videos.
7. A student Rater Practice package of one 'how to' video was developed.
8. A report page showing all RP results was added to the instructor dashboard.

III. Data Collection

For the dispersion analysis we used the peer evaluation data from 804 students from Fall 2015 and 912 students from Fall 2017. For the Social Relation Model [SRM] analysis our

control group was 159 teams of 4 students from Fall 2015 class where RP was required and completed by each member of the team. This provided a SRM control group of 636 students with RP scores who used the RP 2.0 model. For our intervention group we randomly selected 157 teams from a Fall 2017 class with a total of 628 students who used the RP 3.0 model.

IV. Data Analysis

The measure of dispersion used in this analysis is defined as the standard deviation of each student's rating of themselves as well as their teammates across the dimensions being used. To be precise, Figure 2 shows a sample of the raw quantitative peer evaluation data. For a team with 3 members completing a peer evaluation on 4 dimensions (C, I, K & E), each rater in the team contributes to a 3×4 rating matrix; with all 3 raters' response combined column wise forms a 3×12 matrix. The standard deviations for each student were calculated for each row. In Figure 2, this was the standard deviation for each of the three rows under the four "Rater 1" columns. Then the three row-wise standard deviations were averaged, then placed in a matrix of average dispersions and referred to as the dispersion for the Rater 1's ratings for all the team members including Rater 1. This procedure was repeated for Rater 2 and 3 accordingly. The same methods were used to calculate the dispersion matrix for all the teams as well as the FYE control group.

			Rater 1				Rater 2		
Student ID	Team ID	Ratee #	C	I	K	E	C	I	K
A001	1	1	3	4	3	4	3	5	3
A002	1	2	4	3	3	4	3	5	3
A003	1	3	5	5	5	3	5	5	4

Figure 2: Raw data from CATME Peer Evaluation

A One-Way ANOVA statistical analysis was used to compare the differences in dispersions between the RP control and intervention groups. The significance, if any, in dispersion was then calculated using the program R where dispersion of ratings between two different peer evaluations is compared. Holm-Bonferroni step-down adjustment for multiple comparisons was adopted to control family-wise error so that we could obtain more powerful results compared with the original Bonferroni method [23]. The effect sizes (Cohen's d) were calculated using differences in estimates and the pooled standard deviations of the deviation factors. An effect size of 0.2 is considered small, 0.5 is considered medium and 0.8 is considered large [24].

V. Findings

As shown in Table 1, dispersion data are not significantly different for the Fall 15 RP control group vs. the Fall 17 RP 3.0 group. However, the Fall 17 RP 3.0 intervention group has significantly lower mean ratings (mean difference 0.182, $p<0.000$) compared with the Fall 15 control group.

TABLE 1
TWO-SAMPLE T-TEST COMPARING 15 FALL CONTROL AND 17 FALL RP3.0

Intervention Sample (n1=912)	Peer Review Time	Intervention Sample (n2=804)	Peer Review Time	Stdv Difference dispersion	t-stat difference Dispersion	P value* difference Dispersion	Differenc e mean Rating	t-stat difference mean	P value difference mean
15F RP2.0	1	17F RP3.0	1	0.00158	0.13	0.8945	0.1812	7.11	<0.0001
*Satterthwaite method to calculate degree of freedom for unequal variances									

Table 2 shows that in the SRM analysis Rater Variance declined (36% vs. 30%), and Relationship Variance increased (34% to 40%). There was no significant change in Target Variance.

TABLE 2
VARIANCE COMPONENTS FOR RP 2.0 AND RP 3.0

Variance Component	*RP 2.0*			*RP 3.0*		
	Estimate	SE	%	Estimate	SE	%
Rater	0.204	0.021	**35.7**	0.142	0.016	**29.6**
Target	0.176	0.022	**30.7**	0.149	0.021	**30.9**
Relationship	0.192	0.011	**33.5**	0.19	0.011	**39.5**
Actor effect reliability	0.735			0.66		
Target effect reliability	0.704			0.67		

Note: $N_{RP3.0}$ = 157 teams, 628 individuals; $N_{RP\ 2.0}$ = 159 teams, 636 individual

VI. Conclusion

Since CATME Rater Practice was introduced in August, 2017, it has been played by students over 330,000 times and assigned as homework in over 7,000 peer evaluation surveys. The one significant effect from the RP upgrade tested in this analysis was a lower mean score from students playing rater practice before submitting their first peer evaluation. The pattern of rating dispersions remained unchanged in the first peer evaluation. For SRM results we found one positive change, the lowering of Rater Effects. Therefore, in our initial analysis we found preliminary evidence of quality improvements in FYE peer evaluations. A notable limitation of this work is that while this analysis was based on a large sample of students and teams, it was collected from a one cohort of a single course.

VII. Future Research

We intend to continue conducting our analysis on subsequent peer evaluations for the same sample and to gather additional comparative samples. An additional effect of the possible convergence of self and peer ratings will also be examined. These analysis extensions should be available for discussion in Fall 2018.

Acknowledgment

CATME is supported by National Science Foundation Grants 1431694.

REFERENCES

[1] R. Alsop, "Playing well with others," in *The Wall Street Journal*, ed. September, 9, 2002.

[2] Baker M., "Why Teamwork is Important in the Workplace," in *AIB Official Blog*, ed, 2014.

[3] Elrick L. , "The Importance of Teamwork Skills in Work & School," Rasmussen College2015.

[4] Dictionary.com. teamwork [Online]. Teamwork Definition [Online]. Available: http://www.dictionary.com/browse/teamwork?s=t

[5] J. C. Chen and J. Chen, "Testing a new approach for learning teamwork knowledge and skills in technical education," *Journal of Industrial Technology,* vol. 20, no. 2, pp. 37-46, 2004.

[6] R. A. Layton, M. L. Loughry, M. W. Ohland, and G. D. Ricco, "Design and validation of a web-based system for assigning members to teams using instructor-specified criteria," *Advances in Engineering Education,* vol. 2 no. 1, pp. 1-28, 2010.

[7] N. P. Mero and S. J. Motowidlo, "Effects of rater accountability on the accuracy and the favorability of performance ratings," *Journal of Applied Psychology,* vol. 80, no. 4, pp. 517-524, 1995.

[8] N. P. Mero, S. J. Motowidlo, and A. L. Anna, "Effects of Accountability on Rating Behavior and Rater Accuracy," *Journal of Applied Social Psychology,* vol. 33, pp. 2493-2514, 2003.

[9] M. W. Ohland *et al.*, "The Comprehensive Assessment of Team Member Effectiveness: Development of a Behaviorally Anchored Rating Scale for Self and Peer Evaluation," *Academy of Management Learning & Education,* vol. 11, no. 4, pp. 609-630, 2012.

[10] Loughry M., Ohland M., and Moore D., "Development of a Theory-Based Assessment of Team Member Effectiveness," *Educational and Psychological Measurement,* vol. 67, no. 3, pp. 505-524, 2007.

[11] E. J. Thomas, "Improving teamwork in healthcare: current approaches and the path forward," *BMJ quality & safety,* 2011.

[12] M. L. Loughry, M. W. Ohland, and D. J. Woehr, "Assessing Teamwork Skills for Assurance of Learning Using CATME Team Tools," *Journal of Marketing Education,* vol. 36, no. 1, pp. 5-19, 2014.

[13] Ohland M. W. , Layton R. A. , Loughry M. L. , and Yuhasz A. G. , "Effects of behavioral anchors on peer evaluation reliability," *Journal of Engineering Education,* vol. 94, pp. 319-326, 2005.

[14] T. Poling, D. J. Woehr, L. M. Arciniega, and A. Gorman, "The impact of personality and value diversity on team performance.," in *Annual Meeting for the Society for Industrial and Organizational Psychology*, Dallas, TX, 2006.

[15] Bak. Natalia, C. Lally, Yuchen. Cao, and Ferguson D. , "Evaluation of Training in the CATME Peer Evaluation Schema,," presented at the The Purdue Undergraduate Research Symposium, West Lafayette,IN,, 2018.

[16] C. J. Resick, M. W. Dickson, J. K. Mitchelson, L. K. Allison, and M. A. Clark, "Team composition, cognition, and effectiveness: Examining mental model similarity and accuracy," *Group Dynamics: Theory, Research, and Practice,* vol. 14, no. 2, pp. 174-191, 2010.

[17] Cohen J. , *Statistical power analyses for the social sciences*. Hillsdale, NJ: Lawrence Erlbauni Associates, 1988.

[18] M. D. Back and D. A. Kenny, "The social relations model: How to understand dyadic processes," *Social and Personality Psychology Compass,* vol. 4, no. 10, pp. 855-870, 2010.

[19] Loignon Andrew, Woehr David J., Shumski Thomas Jane, Misty Loughry, Ohland Matthew W., and Ferguson Daniel, "Facilitating Peer Evaluation in Team Contexts: The Impact of Frame-of-Reference Rater Training," *Academy of Management Learning & Education,* vol. 16, no. 4, pp. 562-579, 2017.

[20] J. C. Burguillo, "Using game theory and Competition-based Learning to stimulate student motivation and performance.," *COMPUTERS & EDUCATION,* vol. 55, no. 2, pp. 566-575, 2010.

[21] R. Garris, R. Ahlers, and J. E. Driskell, "Gmes, Motivation and Learning: a Research and Practice Model," *SIMULATION & GAMING,* vol. 33, no. 4, pp. 441-467, 2002.

[22] Ferguson Daniel, Lally Chad, Somnooma Hilda, Murch Olivia, and Ohland Matthew W, "Using Frame-of-Reference Training to Improve the Dispersion of Peer Ratings in Teams," in *Frontiers in Education*, Eire PA, 2016.

[23] Holm S., "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics,* vol. 6, no. 2, pp. 65-70, 1979.

[24] J. Cohen, *Statistical power analysis for the behavioral sciences* (no. Second Edition). Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.